DOCUMENT RESUME

ED 413 757 FL 024 778

AUTHOR Hladka, Barbora; Hajic, Jan
TITLE A Simple Czech and English Probabilistic Tagger: A Comparison.
PUB DATE 1995-00-00
NOTE 7p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computational Linguistics; Computer Software; Contrastive Linguistics; *Czech; *Discourse Analysis; English; Foreign Countries; *Language Research; Linguistic Theory; Programming; Statistical Analysis; *Structural Analysis (Linguistics); Uncommonly Taught Languages
IDENTIFIERS *Inflection (Grammar); Language Corpora

ABSTRACT

An experiment compared the tagging of two languages: Czech, a highly inflected language with a high degree of ambiguity, and English. For Czech, the corpus was one gathered in the 1970s at the Czechoslovak Academy of Sciences; for English, it was the Wall Street Journal corpus. Results indicate 81.53 percent accuracy for Czech and 96.83 percent accuracy for English, representing a higher level of accuracy than expected for Czech. Several simple improvements in the Czech tagging system were identified. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 757

A Simple Czech and English Probabilistic Tagger: a Comparison

Barbora Hladká, Jan Hajič

Institute of Formal and Applied Linguistics
Malostranské nám. 25
118 00 Praha 1
Tel.: +42 2 21914 288
Fax: +42 2 21914 309
E-mail: {hladka, hajic}@ufal.mff.cuni.cz

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Norbert Volz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

FL024778

1. Introduction

Highly inflectional languages like Czech pose a special problem for morphology disambiguation (which is usually called tagging). For example, the ending -u is not only highly ambiguous, but at the same time it carries a complex information: it corresponds, e.g., to genitive singular for inanimate nouns, or dative singular for animate nouns, or accusative singular for feminine nouns, or first person singular present tense active participle for certain verbs.

Given the success of statistical methods in different areas including text tagging we wanted to try them even for the Czech language one of the main features of which is a rich inflection displaying a high degree of ambiguity. Originally we expected that the result would be plain negative, getting not more than about two thirds of the tags correct. However, as we show later, we got better results than we had expected.

We used the same statistical approach to tag both the English text and Czech text. For English, we obtained results comparable with the results presented in [Brill 1993] (who uses different methods). For Czech, we obtained results which are less satisfying than those for English results.

2. Data Used

2.1 For Czech

For training, we used the corpus collected at the beginning of the 70ies in the Czechoslovak Academy of Sciences. The corpus was originally hand-tagged, including the lemmatization and syntactic tags. The complete size of the corpus is 600k tokens. We had to do some cleaning and conversion, as we were interested in the words and tags only.

2.2 For English

For training, we used Wall Street Journal [Marcus, Santorini, Marcinkiewicz 1993]. We had to change the format of WSJ to prepare it for our tagging software.

3. Tags

3.1 Czech tags

The original tag system (in the hand-tagged corpus) was too detailed to use it directly. We disregarded all the other information (lemmatization and syntactic tags) from the training data. We used the traditional division into the part of speech tagger classes. Each class contains many tags for each combination of morphological categories. For a description of the tags for the part of speech classes see Table 1. The first letter represents the tag for the part of speech class and it is followed by the morphological categories for the given class. We used special tags for sentence boundaries, punctuation and "unknown tag". We used 1171 different tags in our experiment for Czech. They were manually derived from the training corpus.

nouns	**N**		gender number case	
	abbreviation		**Z**	
adjectives	**A**		gender number case degree negation	
verbs	**V**			
	infinitive		**T**	negation
	transgressive		**W**	number tense voice gender negation
	common			person number voice tense mood gender negation
pronouns	**P**			
	personal		**P**	person number case 3 gender number case
	possessive		**R**	gender-of-the-possessive number-of-the-possessive case person gender number
		svůj	**S**	gender number case
		se	**E**	case
		others	**D**	gender number case negation
adverbs	**O**			
conjunctions	**S**			
numbers	**C**			
prepositions	**R**			
interjections	**F**			
particles	**K**			

Table 1

BEST COPY AVAILABLE

For example:
NMS1 (noun, masculinum animate, singular, nominative)
NNP7 (noun, neuter, plural, instrumental)
VTA (verb, infinitive, affirmative)
V3SAPOMA (verb, 3rd person, singular, active, present tense, indicative, mas. anim., affirmative)
PP2P7 (personal pronoun, 2nd person, plural, instrumental)
AFP32N (adjective, femin. plural, dative, comparative, negative)

3.2 English tags

We used *The Penn Treebank* tagset which contains 36 Part-Of-Speech tags and 12 other tags (for punctuation and the currency symbol). A detailed description is available in [Santorini 1990].

4. The algorithms

We have used Merialdo's methods (described e.g., in [Merialdo 1992]). The tagging procedure selects a sequence of tags T for the sentence W:

$$\Phi : W \rightarrow T = \Phi(W).$$

In this case the optimal tagging procedure is

$$\Phi(W) = \underset{T}{\operatorname{argmax}} Pr(T \mid W) = \underset{T}{\operatorname{argmax}}\, Pr(T \mid W) * Pr(W) = \underset{T}{\operatorname{argmax}} Pr(W \mid T) *$$

$$\underset{T}{\operatorname{argmax}} Pr(W \mid T) * Pr(T)$$

Our implementation is based on generating the (W, T) pairs by a probabilistic model using approximations of probability distributions Pr(W | T) and Pr(T).

The Pr(T) is based on tag bigrams, and Pr(W | T) is approximated as the product of $Pr(w_i \mid t_i)$. The parameters have been estimated by the usual maximum likelihood training method, i. e. we approximated them as the relative frequencies found in the training data, smoothing them accordingly using the unigram frequencies and the uniform distribution.

5. The results

	Experiment for Czech	Experiment for English
corpus	**Czech hand-tagged**	**Wall Street Journal**
trainig data (tokens)	**621 015**	**1 287 749**
trainig data (words)	**72 445**	**51 433**
trainig data (tags)	**1 171**	**45**
training data (the average number of tags per token)	**3,65**	**2,34**
test data (tokens)	**1 294**	**1 294**
incorrect tags	**56**	**41**
tagging accuracy	**81,53%**	**96,83%**

To illustrate the results of our tagging procedures, we present here an example from the tagged test text. The cases of incorrect tag assignment are denoted by boldface letters.

tagged word | hand-assigned tag | result of the tagging programme

Czech test text

jménem | Rjménem | NNS7
úv | NZ | NZ
Ksč | NZ | NZ
pozdravil | V3SAMOMA | NZ
Davisovou | NFS4 | **NZ**
Pavel | NMS1 | NMS1
Auersperg | NMS1 | NMS1
W_SB | T_SB | T_SB
účastníci | NMP1 | NMP1
shromáždění | NNS2 | NNS2

English test text

In | IN | IN
the | DT | DT
lengthy | JJ | JJ
discussion | NN | NN
that | IN | **WDT**
followed | VBD | VBD
, | , | ,
Mr. | NNP | NNP
Buffett | NNP | NNP
said | VBD | VBD
: | : | :

6. Conclusion

The results, however they might seem negative compared to English, are still better than our original expectations. We would like to improve current approach by another simple measures. For example, the average number of tags per token will increase after a morphological analyser is added as the front end to the tagger (serving as the "supplier" of possible tags). We also plan to use trigrams instead of bigrams after we collect more data for Czech. Finally, certain tagset reductions be carried one, as the original tagset (even after the reductions mentioned above) is too detailed (in the sense that it distinguishes tags hardly distinguishable by human annotators). We are also working on independent predictions for certain grammatical categories and the lemma itself, but the final shape of the model has not yet been decided. This would mean to introduce constraints on possible combinations of morphological categories and take them into account when "assembling" the final tag.

References

Brill, E. 1993. "A Corpus Based Approach To Language Learning". Dissertation in Department of Computer and Information, Science, University of Pennsylvania.

Marcus, M. P., B. Santorini and M.A. Marcinkiewicz. 1993. "Building a large annotated corpus of English: the Penn Treebank". To appear in Computational Linguistics. (forthcoming)

Merialdo, B. 1992. "Tagging text with a probabilistic model". Computational Linguistics 20(2), 155–171.

Santorini, B. 1990. "Part of Speech tagging guidelines for the Penn Treebank Project". Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.